Convegno Euro PVM/MPI 2007

Maurizio Cremonesi
CILEA, Segrate

Abstract
La 14esima edizione del congresso EuroPVM/MPI si è svolta a Parigi dal 29 settembre al 3 ottobre 2007. Tra gli argomenti di maggior interesse si è discusso di comunicazioni collettive e *one-side*, estensione ed evoluzione di MPI, *fault tolerance*, studio delle prestazioni, architetture parallele e GRID.

The 14th issue of the EuroPVM/MPI conference was been held in Paris from September 29th to October 3rd 2007. Among the most interesting topics presented were collective and one-side communications, MPI extension and evolution, fault tolerance, performance evaluation, parallel architectures and GRID.

Keywords: Calcolo parallelo; PVM, MPI; Parallelismo massivo.

I congressi EuroPVM/MPI sono un punto di riferimento per la comunità degli sviluppatori di calcolo parallelo in PVM e MPI; si può considerare senz'altro l'evento come di elezione per la discussione di tutte le problematiche inerenti lo sviluppo di programmi efficienti su piattaforme parallele e l'occasione per l'incontro tra sviluppatori e utilizzatori di MPI. I congressi EuroPVM/MPI sono organizzati ogni anno in città europee diverse; la prima edizione è stata organizzata a Roma nel 1994; le ultime edizioni si sono svolte a Bonn (2006) e Sorrento (2005).

EuroPVM/MPI 2007 [1] è stato organizzato a Parigi dall'Istituto Nazionale francese per la Ricerca nell'Informatica e nell'Automazione (INRIA), ha visto la partecipazione di numerosi relatori e partecipanti provenienti soprattutto dall'Europa, ma anche un po' da tutto il mondo, dall'Asia e dalle Americhe e ha avuto il sostegno di diversi sponsor importanti.

Al congresso del 2007 gli organizzatori hanno invitato sei ricercatori di fama mondiale per presentare gli sviluppi più recenti del calcolo parallelo: Tony Hey, co-autore della prima versione di MPI, AlGeist, co-autore di PVM, Satoshi Matsuoka, pioniere del calcolo in GRID, Ewing Lusk, uno degli sviluppatori di MPICH, George Bosilca, uno degli sviluppatori di OpenMPI; Bernd Mohr, pioniere dei programmi per l'analisi delle prestazioni.

Il congresso è stato preceduto da tre tutorial sull'utilizzo delle funzioni MPI più avanzate e sul *debug* di programmi MPI. La qualità dei lavori presentati è stata assicurata da un comitato di revisori costituito dai principali sviluppatori e utilizzatori di MPI, sia in Europa che in America.

Fig. 1 – Il logo di EuroPVM/MPI 2007 è stato disegnato da Ala Rezmerita

PVM ancora tra noi

Si deve segnalare che al congresso di quest'anno l'Italia non ha brillato per partecipazione: tra i relatori c'era solo la MBDA Italia SpA, che ha presentato un'applicazione per il controllo di un'apparecchiatura industriale in tempo reale. L'applicazione si basa sull'utilizzo di Harness, un *middleware framework* adattabile, basato su *plugin*, adatto allo sviluppo di programmi paralleli che non richiedono grosse risorse di calcolo. Harness si basa su software di comunicazione evoluto da PVM. L'unico altro ente italiano presente, ma solo come uditore, è stato il CILEA.

Se a Roma usano un gestore del parallelismo distribuito basato sui concetti di PVM perché snello e di poche pretese, in Turchia PVM è

utilizzato in applicazioni *data-mining* su piccoli *cluster* con processori a basse prestazioni. Nessun altro ha presentato lavori su PVM a questo congresso.

Fig. 2 – EuroPVM/MPI 2007 è stata ospitata nel Centre Espace Etoile Saint Honore in Via Balzac

La fine del calcolo seriale

Uno degli argomenti più discussi a Euro PVM/MPI 2007 è stato l'impatto dell'imminente rivoluzione architetturale nelle piattaforme di calcolo ad alte prestazioni. Questo preoccupa gli sviluppatori di algoritmi di calcolo parallelo e degli strumenti per lo sviluppo del software; tuttavia rimane ancora qualche anno per prepararsi, prima che gli effetti delle nuove tecnologie si facciano sentire. I costruttori di processori negli ultimi anni si sono scontrati con il problema di non poter più aumentare la velocità di calcolo riducendo semplicemente le dimensioni dei circuiti elettronici, perché all'aumento della frequenza corrisponde ormai una crescita sproporzionata dei consumi di energia, con la conseguente impossibilità di raffreddare adeguatamente ed economicamente il processore. Si prevede che la Legge di Moore resterà valida ancora per almeno un decennio; tuttavia questo non dipenderà dall'aumento della velocità del processore ma dall'incremento del numero di unità di calcolo nel singolo chip, aumentando così il parallelismo della macchina. Se negli ultimi decenni si poteva contare nel raddoppio della velocità del processore ogni circa 18 mesi, da ora si può solo contare nell'aumento del parallelismo. Se attualmente sono in produzione nodi con 4 core, si può pensare che tra qualche anno si potranno costruire nodi con almeno duecento core, probabilmente molti di più. Come conseguenza entro alcuni anni saranno operative diverse piattaforme di calcolo con un numero spropositato, per gli standard attuali, di unità di calcolo. Il Dipartimento per l'Energia degli USA (DOE) e il National Science Fundation (NSF) hanno in programma lo sviluppo di piattaforme di calcolo Cray, che porterebbero alla realizzazione di una macchina da 1 PF nel 2009 e di una macchina da 20 PF intorno al 2011. In particolare la macchina da 20 PF avrebbe nome Cray Cascade [2] e potrebbe avere 800.000 core distribuiti in 6.224 nodi, più di 100 core per nodo; ogni nodo disporrebbe di una RAM da 256 GB per una memoria totale di 1,5 PB; la capacità disco globale sarebbe invece di 46 PB. La macchina sarebbe costituita da 264 *cabinet* e occuperebbe un'area di circa 800 m^2, assorbendo una potenza complessiva di 15 MW. In contrasto la macchina da 1 PF, in servizio forse entro 2 anni, avrebbe "solo" 98.304 core distribuiti in 24.576 nodi, 4 core per nodo, con 175 TB di RAM totale. Queste macchine saranno destinate alla soluzione di problemi sull'evoluzione del clima planetario, la fusione nucleare, i nanomateriali, lo studio dei sistemi molecolari in cellule e batteri.

Fig. 3 – Il Centre Espace Etoile Saint Honore popolato dai partecipanti alla conferenza

I cluster del futuro non saranno solamente massicciamente paralleli ma molto probabilmente avranno unità di calcolo eterogenee. L'eterogeneità è già una realtà in cluster come il Roadrunner dei Los Alamos National Laboratoires (LANL) e il giapponese TSUBAME.

Roadrunner è un cluster costituito da processori AMD Opteron e IBM Cell BE; questa macchina [3] è annunciata come la prima piat-

taforma di calcolo in grado di raggiungere 1 PF di potenza entro il 2008.

Viceversa, il cluster TSUBAME è stato installato nel 2006 presso l'Istituto di tecnologia di Tokyo e a fine 2007 avrà un centinaio di TF di potenza di picco, 22 TB di RAM e 1,6 PB di spazio disco. La macchina occupa circa 350 m^2 in tre sale diverse, con comunicazioni tra una sala e l'altra realizzate in fibra ottica; all'interno di ogni sala i collegamenti sono in fili di rame. Il consumo di energia è di meno di 1 MW a piena potenza. Circa 2/5 della potenza di calcolo sono forniti dagli acceleratori Clearspeed PCI-X; la capacità di calcolo potrebbe essere aumentata accoppiando un acceleratore a ogni nodo AMD Opteron, ma l'utilizzabilità degli acceleratori dovrebbe essere migliorata, con nuovi algoritmi e tecniche software. Si ritiene che la giusta combinazione di processori *general-purpose* e acceleratori specializzati permetterà di costruire calcolatori di successo e ad alta efficienza.

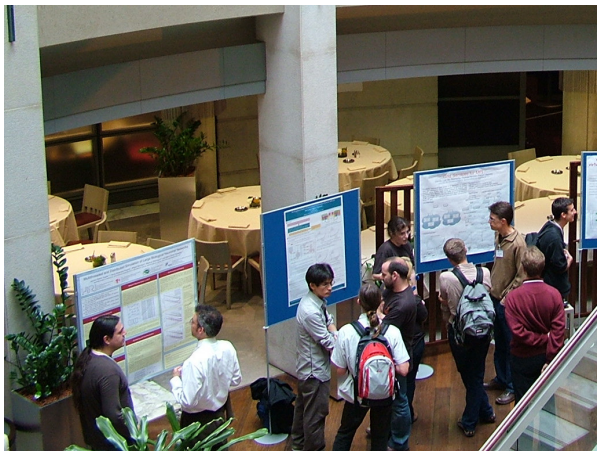

Fig. 4 – Il convegno EuroPVM/MPI 2007 ha visto la partecipazione di numerosi giovani

Gestire l'iper parallelismo

Se il Fortran e il C, insieme a MPI, dominano ancora il software nel campo della programmazione ad alte prestazioni, non si può prevedere quali strumenti verranno usati per macchine del parallelismo di milioni di core.
Nel 2002 negli USA la Defense Advanced Research Projects Agency (DARPA) ha iniziato il progetto "High Productivity Computing Systems" [4], allo scopo di migliorare la programmabilità delle macchine ad altissimo parallelismo. La Cray ha attualmente in sviluppo il linguaggio di programmazione Chapel, ispirato tra l'altro all'HPF (High Performance Fortran) che nelle intenzioni dovrebbe migliorare la programmabilità di macchine a elevato parallelismo come il Cray Cascade.

I suggerimenti che attualmente vengono dati per migliorare le prestazioni su macchine iper parallele sono: migliorare le funzioni collettive, aggregare le comunicazioni, migliorare l'efficienza dell'I/O con una gestione più intelligente e coordinata. Questo può essere realizzato sia agendo sul programma di calcolo, ma forse più efficacemente agendo sulle librerie di gestione dei processi come MPI, sia strutturando le applicazioni a livelli diversi di complessità organizzativa. D'altra parte, se ogni nodo di un cluster contiene diverse decine di processi indipendenti e ha pochi canali di comunicazione con gli altri nodi della macchina, non sarebbe fuori luogo pensare di realizzare le applicazioni pur all'interno della singola piattaforma di calcolo con tecniche di tipo GRID computing.

Gli sviluppatori si interrogano sul futuro di MPI e si chiedono se sopravviverà. Il problema non riguarda tanto la quantità di nodi, perché programmi MPI sono stati realizzati con successo su macchine da 100.000 nodi, anche se nessuno può prevedere cosa succederà quando i nodi saranno 10 o 100 volte di più. Un grosso problema è la continuità delle esecuzioni [5]. Macchine con milioni di nodi sono particolarmente sensibili a qualunque disturbo e oltretutto sarebbe statisticamente improbabile che su centinaia di migliaia di processi attivi nessuno subisca disturbi più o meno gravi in diversi giorni di attività continuativa. Attualmente, la soluzione più diffusa per applicazioni MPI in caso di interruzione imprevista di un processo è il *restart* dell'intera esecuzione, preferibilmente partendo dal check-point più recente. Ma non è pratico né desiderabile interrompere un'esecuzione che impiega milioni di processi per il fallimento di un singolo processo; oltretutto il check-pointing in tali sistemi potrebbe addirittura peggiorare la stabilità del sistema stressando oltremodo l'I/O.
Un altro grosso problema riguarda l'affidabilità dei programmi. Il tool di debug Apprentice della Cray è stato sperimentato su 11.000 nodi, ma sarebbe ancora possibile usarlo con efficacia su 100.000 o più nodi? Forse la continuità e l'affidabilità di esecuzioni iper parallele possono essere realizzate utilizzando la ridondanza: facendo girare lo stesso lavoro su macchine diverse contemporaneamente e prevedendo punti di sincronizzazione durante i quali si con-

frontano i risultati parziali sarebbe probabilmente possibile salvare un'esecuzione anche quando alcuni processi vengono meno.

Qualunque siano i problemi da affrontare, qualunque sia il futuro del calcolo, è compito di eventi come Euro PVM/MPI saper riunire i migliori ricercatori a livello mondiale allo scopo di far nascere e sviluppare le idee più fruttuose e le soluzioni più originali [6].

Bibliografia

[1] URL: http://pvmmpi07.lri.fr/
[2] URL: http://www.cray.com/products/programs/cascade.html
[3] URL: http://www.lanl.gov/orgs/hpc/roadrunner/index.shtml
[4] "Languages for high-productivity computing: the DARPA HPCS language project", Ewing Lusk, Katherin Yelick
[5] "Development of Naturally Fault Tolerant Algorithms for Computing on 100,000 Processors", Al Geist, Christian Engelmann, Oak Ridge National Laboratory, URL: www.csm.ornl.gov/~geist
[6] LNCS 4757, "Recent Advances in Parallel Virtual Machine and Message Passing Interface", Franck Cappello, Thomas Herault, Jack Dongarra Eds., Springer, 2007.